ELSEVIER Stochastic Processes and their Applications 55 (1995) 57–64

stochastic processes and their applications

Generalised two-sample U-statistics and a two-species reaction-diffusion model

Mathew D. Penrose

Department of Mathematical Sciences, University of Durham, South Road, Durham DH1 3LE, UK

Received 4 July 1993; revised 23 May 1994

Abstract

Consider a random variable of the form $U = \sum f(X_i, Y_j)$, where the sum is over all pairs from independent samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ from two (possibly different) distributions, and f is a given function which may depend on n. We discuss possible limits for the distribution of U when n becomes large with $n/(m+n)$ approaching a fixed limit. We discuss an application to a Brownian motion for the irreversible two-species, diffusion-controlled chemical reaction.

Keywords: U-statistics; Limit laws; Infinitely divisible distributions; Chemistry; Reaction; Diffusion; Brownian motion; Poisson process

1. Introduction

Let $(E, \mathcal{E})$ be a measurable space. On a probability space $(\Omega, \mathcal{F}, P)$ let $X_1, X_2, \ldots,$ be independent indentically distributed (iid) E-valued random variables, and let $Y_1, Y_2, \ldots,$ be iid E-valued variables, independent of $\{X_i, i \geqslant 1\}$. Let $f_n : E \times E \to \mathbb{R}$ be a jointly measurable function, and set

$$U_{n,m} = \sum_{i \leqslant n} \sum_{j \leqslant m} T_{ij}^n, \quad \text{where } T_{ij}^n = f_n(X_i, Y_j).$$

A random variable of the form of $U_{n,m}$ is a *two-sample U-statistic.* When the function f_n does not depend on n, the asymptotic theory for $U_{n,m}$ is well-known (Lehmann, 1951). For example, if $E_1 = E_2 = \mathbb{R}$ and $f(x,y) = \mathbf{1}_{\{x>y\}}$ (here $\mathbf{1}_{\{\}}$ denotes indicator function), then $U_{n,m}$ is the Wilcoxon–Mann–Whitney statistic. For more information on U-statistics, see Lee (1990) and Randles and Wolfe (1979).

Here we consider the asymptotic theory when f_n varies with n, as did Jammalamadaka and Janson (1986) in the one-sample case. As in that paper, we obtain a

SSDI 0304-4149(94)00039-V

range of infinitely divisible laws as possible limiting distributions for $U_{n,m}$ when n and m become large in a linked manner.

Theorem 1. Suppose that (m_n) is a sequence of integers such that

$$n/(n+m_n) \to p \in (0,1), \quad \textit{as } n \to \infty.$$

Suppose

$$\lim_{n\to\infty} n^3 E[T_{11}^n T_{12}^n] = \lim_{n\to\infty} n^3 E[T_{11}^n T_{21}^n] = 0. \tag{1}$$

Suppose also that there are real numbers a_n and a finite positive measure ψ on $\mathbb{R}$, such that if F_n denotes the distribution function of T_{11}^n, then

$$nm_n(t^2/(1+t^2))\,\mathrm{d}F_n(t) \to_c \mathrm{d}\psi(t), \quad \textit{as } n \to \infty$$

and

$$nm_n E\left[\frac{T_{11}^n}{1+(T_{11}^n)^2}\right] - a_n \to 0, \quad \textit{as } n \to \infty,$$

where $\to_c$ denotes complete convergence (pointwise convergence on the continuity set of ψ, and at $\pm\infty$). Then

$$U_{n,m_n} - a_n \to V, \quad \textit{in distribution as } n \to \infty,$$

where V has the infinitely divisible characteristic function

$$E\,\mathrm{e}^{\mathrm{i}Vt} = \exp \int_{-\infty}^{\infty} \left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right)\left(\frac{1+x^2}{x^2}\right) \mathrm{d}\psi(x).$$

Of special interest is the case where f_n is a zero-one function.

Corollary 2. Suppose f_n is zero-one, and set $q_n = ET_{11}^n$. Suppose the sequence (m_n) is as in Theorem 1. Then

(i) if (1) holds, and $nm_nq_n \to \lambda > 0$, then $U_{n,m_n} \to \text{Poisson}(\lambda)$ in law;

(ii) if $n^2 q_n \to \infty$, and

$$\lim_{n\to\infty} (n/q_n) E[T_{11}^n T_{12}^n] = \lim_{n\to\infty} (n/q_n) E[T_{11}^n T_{21}^n] = 0, \tag{2}$$

then $nq_n \to 0$ and

$$\left(U_{n,m_n} - EU_{n,m_n}\right)/(\text{Var}\, U_{n,m_n})^{1/2} \to \mathrm{N}(0,1) \quad \textit{in law}. \tag{3}$$

Proof of (ii). Since $E[T_{11}^n T_{12}^n] = E(E[T_{11}^n | X_1]^2) \geqslant q_n^2$ by Cauchy–Schwarz, the first conclusion $nq_n \to 0$ follows from (2). It is easy to show that $\text{Var}\, U_{n,m_n} = nm_nq_n(1+\mathrm{o}(1))$, and then to deduce (3) from Theorem 1. □

For one possible application, let E be the space $\mathbb{R}^d$, equipped with the Euclidean norm $|\cdot|$. Suppose $X_i, 1 \leqslant i \leqslant n$, are independent observations from a distribution with density $f(x), x \in \mathbb{R}^d$, and $Y_j, 1 \leqslant j \leqslant m$, are observations from a density $g(y), y \in \mathbb{R}^d$. If we set $f_n(x,y) = \mathbf{1}_{\{|x-y| \leqslant r_n\}}$, where (r_n) is a suitable sequence converging to zero, then $U_{n,m}$ counts the number of pairs of observations, one from each sample, which are close. This statistic can be the basis for a non-parametric test for equality of location parameters in multidimensional Euclidean space. Indeed, suppose for some known density h that

$$f(x) = h(x + \theta_1), \qquad g(x) = h(x + \theta_2), \quad x \in \mathbb{R}^d,$$

and we wish to determine whether $\theta_1 = \theta_2$. Then Corollary 2 can be used to show that U_{n,m_n} has a limiting distribution that is Poisson or Gaussian, according to the choice of (r_n), with mean maximised by setting $\theta_1 = \theta_2$.

In Section 3 we shall present a further application of the limit theory, to a model for chemical reaction with two types of molecules, executing Brownian motions in a suspension fluid, with molecules of opposite type annihilating one another whenever they collide. In this case, E is a space of continuous functions.

2. Proof of Theorem 1

Set $m = m_n$, and $U_n = U_{n,m_n}$. Following Jammalamadaka and Jansen (1986, p. 1353), let M, N be Poisson random variables, independent of one another and of $\{X_i, Y_i, i \geqslant 1\}$, with means n and m respectively. Set

$$U'_n = \sum_{i=1}^{N} \sum_{j=1}^{M} T^n_{ij}.$$

Then if $N \geqslant n, M \geqslant m$, we have

$$U'_n - U_n = \sum_{n<i\leqslant N} \sum_{1\leqslant j\leqslant m} T^n_{ij} + \sum_{1\leqslant i\leqslant N} \sum_{m<j\leqslant M} T^n_{ij}.$$

If $N \geqslant n, M < m$, then

$$U'_n - U_n = \sum_{n<i\leqslant N} \sum_{1\leqslant j\leqslant m} T^n_{ij} - \sum_{1\leqslant i\leqslant N} \sum_{M<j\leqslant m} T^n_{ij},$$

and there are similar expressions in the other cases ($N < n, M \geqslant m$ and $N < n, M < m$). Hence

$$E[(|U'_n - U_n|) \,|\, N, M] \leqslant (|N-n|m + N|M-m|)E[|T^n_{11}|].$$

By Cauchy–Schwarz, $E|N-n| \leqslant n^{1/2}$. Similarly for M, so for some $c > 0$,

$$E(|N-n|m + N|M-m|) \leqslant cn^{3/2}.$$

By (1),

$$n^{3/2}E|T^n_{11}| \leqslant n^{3/2}(E(E[T^n_{11}|X_1]^2))^{1/2} \to 0,$$

so that

$$E[|U'_n - U_n|] \to 0 \quad \text{as } n \to \infty.$$

Thus it suffices to prove the desired convergence in law for U'_n instead of U_n.

Let $p_n = (n/(n+m))$, so $p_n \to p$ as $n \to \infty$. For each n let $J_1^n, J_2^n, \ldots$, be iid random variables with $P[J_i^n = 1] = p_n$ and $P[J_i^n = 2] = 1 - p_n$, independent of $\{X_i, Y_i, i \geqslant 1\}$. Let

$$Z_i^n = X_i, \quad \text{if } J_n^i = 1; \qquad Z_i^n = Y_i, \quad \text{if } J_n^i = 2.$$

Let $E' = E \times \{1, 2\}$. Let W_i^n denote the pair (Z_n^i, J_n^i). Then $(W_1^n, W_2^n, W_3^n, \ldots)$ is an iid sequence in E'. The law of Z_1^n is a mixture of those of X_1 and Y_1. Define $g_n : E' \times E' \to \mathbb{R}$ by

$$\begin{aligned} g_n((x,i),(y,j)) &= f_n(x,y), && \text{if } i = 1, j = 2, \\ g_n((x,i),(y,j)) &= f_n(y,x), && \text{if } i = 2, j = 1, \\ g_n((x,i),(y,j)) &= 0, && \text{otherwise.} \end{aligned}$$

Let $\overline{N}$ be a Poisson random variable with mean $n + m$. By the decomposability of the Poisson distribution, if we set

$$N' = \sum_{i=1}^{\overline{N}} \mathbf{1}_{\{J_i = 1\}} \quad \text{and} \quad M' = \sum_{i=1}^{\overline{N}} \mathbf{1}_{\{J_i = 2\}},$$

then $(N', M') =^d (N, M)$, where $=^d$ means equality of distribution. Hence,

$$U'_N =^d \sum \sum_{i < j \leqslant \overline{N}} g_n(W_i^n, W_j^n).$$

Then

$$n^3 E|g_n(W_1^n, W_2^n) g_n(W_1^n, W_3^n)| = n^3 p_n (1-p_n)^2 E T_{12}^n T_{13}^n + n^3 p_n^2 (1-p_n) E T_{21}^n T_{31}^n,$$

which converges to 0 by (1). Also, if F'_n is the distribution function of $g_n(W_1^n, W_2^n)$, and F_0 is that of a point mass at zero, then

$$F'_n = 2p_n(1-p_n)F_n + (1 - 2p_n(1-p_n))F_0.$$

If we had $p_n = p$ for all n, the desired infinitely divisible limit for the law of U'_n would now be immediate from (2), (3) and Jammalamadaka and Janson (1986, Theorem 3.2). However, the proof of that result remains valid when the p_n vary, but converge to p as $n \to \infty$, and we are done.

3. A two-species reaction-diffusion model

The following is a model for the irreversible diffusion-controlled chemical reaction $A + B \to \emptyset$. Molecules of type A and B execute Brownian motion in a suspension fluid, and whenever a molecule of type A approaches closely enough to a molecule of type B, they annihilate one another. We shall use Theorem 1 to look at the limiting law of the point process obtained by recording the time and position of each annihilation, when the number of particles becomes large and the radius of interaction becomes small in a linked manner.

This is a two-species version of a model which was studied in the one-species case by Sznitman (1987) and Penrose (1994). A lattice model for the two-species reaction appears in Bramson and Lebowitz (1988, 1991). See references there, and in Clifford et al. (1987) for related models.

Let $d \geqslant 2$. On a probability space $(\Omega_n, \mathcal{F}_n, P_n)$ (with expectation E_n), let $X_1, X_2, \ldots X_n$ be iid Brownian motions in $\mathbb{R}^d$ with diffusion coefficient D_X (i.e. $\mathrm{Var}(X_i(1) - X_i(0)) = D_X \mathrm{Id}$) with initial distribution

$$P[X_i(0) \in \mathrm{d}x] = u_X(x)\,\mathrm{d}x,$$

and let $Y_1, \ldots, Y_{m_n}$ be iid Brownian motions in $\mathbb{R}^d$, independent of the X_i's, with diffusion coefficient D_Y and initial distribution $P[Y_j(0) \in \mathrm{d}y] = u_Y(y)\,\mathrm{d}y$, where u_X and u_Y are bounded probability density functions. Define

$$B_{ij}(t) = (D_X + D_Y)^{-1/2}(X_i(t) - Y_j(t))$$

and

$$\widetilde{B}_{ij}(t) = (D_X + D_Y)^{-1/2}\left(\left(\frac{D_Y}{D_X}\right)^{1/2} X_i(t) + \left(\frac{D_X}{D_Y}\right)^{1/2} Y_j(t)\right).$$

Then $B_{ij}(\cdot)$ and $\widetilde{B}_{ij}(\cdot)$ are independent standard Brownian motions. Let $(r_n, n \geqslant 1)$ be a sequence converging to zero, and let $s_n = (S_d(r_n))^{-1}$, where we set

$$S_d(x) = \log(1/x), \quad d = 2; \qquad S_d(x) = x^{2-d}, \quad d \geqslant 3.$$

Now define

$$T_{ij} = \inf\{t \geqslant 0 : |B_{ij}(t)| \leqslant r_n\}.$$

Thus T_{ij} (which also depends on n) is the first time the particles with motions described by X_i and Y_j come within a fixed distance of one another. Suppose that at time T_{ij} these two particles annihilate one another, provided neither of them has previously been annihilated. Let T^k denote the kth time at which an annihilation takes place: then $0 \leqslant T^1 < T^2 < T^3 < \cdots < T^L$, where L denotes the total number of such times. Thus $\{T^k : 1 \leqslant k \leqslant L\}$ is a subset of $\{T_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m_n\}$. Indeed, for $k \leqslant L$ and $T^k > 0$, the annihilations at time T^k involve exactly 1 particle of each type, say the $i(k)$th particle of type A and the $j(k)$th particle of type B. So $T^k = T_{i(k),j(k)}$. Set

$$Z^k = \widetilde{B}_{i(k),j(k)}(T^k).$$

When n is large, r_n is small and the vector Z^k is close to the vector

$$(D_X + D_Y)^{-1/2}\left(\left(\frac{D_Y}{D_X}\right)^{1/2} + \left(\frac{D_X}{D_Y}\right)^{1/2}\right)\left(X_{i(k)}(T^k) + Y_{j(k)}(T^k)\right)/2.$$

This is a fixed multiple of the vector $\frac{1}{2}(X_{i(k)}(T^k)+Y_{j(k)}(T^k))$, which may be viewed as the position at which collision occurs.

Define the point process (random point measure) η_n in $\mathbb{R}_+ \times \mathbb{R}^d$ to consist of the points $(T^k, Z^k), 1 \leqslant k \leqslant L$, with $T^k > 0$. That is, for any test function f on $\mathbb{R}_+ \times \mathbb{R}^d$, set

$$\eta_n(f) = \sum_{k=1}^{L} f(T^k, Z^k) \mathbf{1}_{\{T^k>0\}}.$$

Thus η_n records the time and place of each annihilation. We endow the space of point measures with the vague topology ($a_n \to a$ if and only if $a_n(f) \to a(f)$, $f \in C_0(\mathbb{R}_+ \times \mathbb{R}^d)$). Our result here is a Poisson weak limit for η_n. To describe it, we need some more definitions.

Let $v(z,\tilde{z})$ $(z,\tilde{z} \in \mathbb{R}^d)$ be the joint density of $(B_{11}(0), \widetilde{B}_{11}(0))$:

$$v(z,\tilde{z}) = (D_X D_Y)^{d/2} u_X \left(\left(\frac{D_X}{D_X+D_Y} \right)^{1/2} (D_X^{1/2} z + D_Y^{1/2} \tilde{z}) \right)$$
$$\times u_Y \left(\left(\frac{D_Y}{D_X+D_Y} \right)^{1/2} (-D_Y^{1/2} z + D_X^{1/2} \tilde{z}) \right).$$

Let $p_t(\cdot)$ denote the Brownian transition density $p_t(x) = (2\pi t)^{-d/2} \exp(-|x|/(2t)), t > 0$, $x \in \mathbb{R}^d$. Let $\pi_d = \pi^{d/2} \Gamma((d/2)+1)$, the volume of the unit ball in $\mathbb{R}^d$. Define

$$C_d = \pi, \quad d=2; \qquad C_d = ((d/2)-1)\, d\pi_d, \quad d \geqslant 3.$$

Define $\lambda(t,x)$ to be C_d times the density of $(B_{11}(t), \widetilde{B}_{11}(t))$ at $(0,x)$:

$$\lambda(t,x) = C_d \int_{R^d} \int_{R^d} p_t(x-\tilde{y}) p_t(y) v(y,\tilde{y})\, dy\, d\tilde{y}, \quad (t>0, x \in \mathbb{R}^d).$$

Theorem 3. Suppose $n/(n+m_n) \to p \in (0,1)$ and $n m_n s_n \to \gamma \in (0,\infty)$ as $n \to \infty$. Then η_n converges in law to a Poisson process with mean measure $\gamma\lambda(t,x)\, dx\, dt$.

In Penrose (1994) we obtained an analogous result for a system with only one type of particle. We also obtained Gaussian limits in the regime $n^2 s_n \to \infty$, $n^{1+\epsilon} s_n \to 0$ for some $\epsilon > 0$. It is possible to derive similar results in the two-species setting, but we shall not do so here.

To prove Theorem 3 we shall need two lemmas. For $f \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$, set

$$\phi_n(f) = \sum_{i=1}^{n} \sum_{j=1}^{m_n} f(T_{ij}, \tilde{B}_{ij}(T_{ij}))$$

with the interpretation that $f(T_{ij}, \widetilde{B}(T_{ij})) = 0$ if $T_{ij} = \infty$. The point process ϕ_n is a close approximation to η_n.

Lemma 4. Suppose $f \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$, with support$(f) \subset (0,\tau) \times \mathbb{R}^d$, some $\tau > 0$. Then under the hypothesis of Theorem 3,

$$P_n[\phi_n(f) = \eta_n(f)] \to 1, \quad as\ n \to \infty.$$

Proof. Since $\phi_n(f)$ differs from $\eta_n(f)$ only as a result of those particles which collide with two or more particles of opposite type before time τ,

$$|\phi_n(f) - \eta_n(f)| \leqslant \|f\|_\infty \left(\sum_i \sum_j \sum_{j' \neq j} \mathbf{1}_{\{T_{ij} \leqslant T_{ij'} \leqslant \tau\}} + \sum_i \sum_{i' \neq i} \sum_j \mathbf{1}_{\{T_{ij} \leqslant T_{i'j} \leqslant \tau\}} \right)$$

We estimate this using the following consequence of Le Gall (1986, Lemma 2.1):

$$\int_{R^d} P[T_{ij} \leqslant \tau | B_{ij}(0) = x]\, \mathrm{d}x \leqslant \text{const.} \times s_n. \tag{4}$$

For $t > 0$, we have

$$P[Y_{j'}(T_{ij}) \in \mathrm{d}y | T_{ij} = t, X_i(T_{ij}) = x] = \left(\int u_Y(z) p_{tD_Y}(y - z)\, \mathrm{d}z \right) \mathrm{d}y.$$

Since u_Y is bounded and $\int p_{tD_Y}(y - z)\, \mathrm{d}z = 1$, this shows that given T_{ij} and $X_i(T_{ij})$, the conditional density of $Y_{j'}(T_{ij})$ is uniformly bounded; hence so is that of $B_{ij'}(T_{ij})$, by c_1 say. Using the strong Markov property at time T_{ij}, and (4) twice, we have for suitable constants c_2 and c_3,

$$\begin{aligned} P[T_{ij} \leqslant T_{ij'} \leqslant \tau] &\leqslant \int_{t=0}^{\tau} P[T_{ij} \in \mathrm{d}t] \int_{R^d} c_1\, \mathrm{d}x P[T_{ij'} \leqslant \tau - t | B_{ij'}(0) = x] \\ &\leqslant c_2 s_n P[T_{ij} \leqslant \tau] \leqslant c_3 s_n^2, \end{aligned} \tag{5}$$

and similarly for $P[T_{ij} \leqslant T_{i'j} \leqslant \tau]$. The result follows since $n^3 s_n^2 \to 0$ here.

Since $B_{ij}(\cdot)$ is standard Brownian motion, the following lemma can be verified by the same argument as in Lemma 2 of Penrose (1994).

Lemma 5. Suppose $h \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$. Suppose for some $\tau > 0$ that for all $x \in \mathbb{R}^d$, $h(\cdot, x)$ is piecewise continuous and $h(t, x) = 0$ for $t > \tau$. Then

$$\lim_{n \to \infty} s_n^{-1} E_n[h(T_{11}, \tilde{B}_{11}(0))] = C_d \int_{R^d} \int_{R^d} \int_0^\tau h(t, \tilde{y}) p_t(y) v(y, \tilde{y})\, \mathrm{d}t\, \mathrm{d}y\, \mathrm{d}\tilde{y} < \infty.$$

Proof of Theorem 3. Suppose $Q \subset \mathbb{R}_+ \times \mathbb{R}^d$ is a finite union of sets of the form $J \times A$, where $J \subset \mathbb{R}_+$ is a bounded interval and $A \subset \mathbb{R}^d$ is a Borel set. Suppose f is the indicator function of Q. Then

$$\phi_n(f) = \sum_{i=1}^{n} \sum_{j=1}^{m_n} W_{ij},$$

where

$$W_{ij} = f(T_{ij}, \widetilde{B}_{ij}(T_{ij})) \mathbf{1}_{\{0 < T_{ij} \leqslant \tau\}}.$$

As in the proof of Theorem 1 of Penrose (1994), we have

$$\lim_{n\to\infty} s_n^{-1} E_n(W_{11}) = \int_{R^d} \int_0^T f(t,x)\lambda(t,x)\,\mathrm{d}t\,\mathrm{d}x = \int\int_Q \lambda(t,x)\,\mathrm{d}t\,\mathrm{d}x.$$

Apply Corollary 2, setting $E = C([0,\infty), \mathbb{R}^d)$ and using (5) to verify that (1) holds here. This result shows $\phi_n(f)$ converges in law to a Poisson distribution with mean $\gamma \int\int_Q \lambda(t,x)\,\mathrm{d}t\,\mathrm{d}x$. By Lemma 1, $\eta_n(f)$ conveges to the same distribution, and convergence to a Poisson process follows from Kallenberg (1973), Theorem 2.3.

References

M. Bramson and J.L. Lebowitz, Asymptotic behavior of densities in diffusion-dominated annihilation reactions, Phys. Rev. Letters 61 (1988) 2397–2400.

M. Bramson and J.L. Lebowitz, Asymptotic behavior of densities for two-particle annihilating random walks, J. Statist. Phys. 62 (1991) 297–372.

P. Clifford, N.J.B. Green and M.J. Pilling, Statistical models of chemical kinetics in liquids, J. R. Statist. Soc. B49 (1987) 266–300.

S.R. Jammalamadaka and S. Janson, Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances, Ann. Probab. 14 (1986) 1347–1358.

O.Kallenberg, Characterization and convergence of random measures and point processes, Z. Wahrsch. Verw. Gebeite 27 (1973) 9–21.

A.J. Lee, U-Statistics (Marcel Dekker, New York, 1990).

J.-F. Le Gall, Sur la saucisse de Wiener et les points multiples du mouvement brownien, Ann. Probab. 14 (1986) 1219–1244.

E.L. Lehmann, Consistency and unbiasedness of certain nonparametric tests, Ann. Math. Statist. 22 (1951) 165–179.

M.D. Penrose, Fluctuations for annihilations of Brownian spheres, Stochastic Process. Appl. 51 (1994) 167–189.

R.H. Randles and D.A. Wolfe, Introduction to the Theory of Non-parametric Statistics (Wiley, New York, 1979).

A.S. Sznitman, Propagation of chaos for a system of annihilating Brownian spheres, Comm. Pure Appl. Math. 40 (1987) 663–690.